Express Mail Label No. EV225207050US
Date of Deposit: September 8, 2003

PROSODIC MIMIC METHOD AND APPARATUS

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims benefit under 35 U.S.C. §119(e) of prior U.S. provisional patent application 60/442,267, entitled "Prosodic Mimic for Commands or Names," filed on January 24, 2003, which is incorporated herein by reference.

TECHNICAL FIELD

The present invention relates to voice-enabled communication systems.

BACKGROUND

Many mobile telephones (here meant to encompass at least data processing and communication devices that carry out telephony or voice communication functions) are provided with voice-assisted interface features that enable a user to access a function by speaking an expression to invoke the function. A familiar example is voice dialing, whereby a user speaks a name or other pre-stored expression into the telephone and the telephone responds by dialing the number associated with that name.

To verify that the number to be dialed or the function to be invoked is indeed the one intended by the user, a mobile telephone can display a confirmation message to the user, allowing the user to proceed if correct, or to abort the function if incorrect. Audible and/or visual user interfaces exist for interacting with mobile telephone devices. Audible confirmations and user interfaces allow a more hands-free operation compared to visual confirmations and interfaces, such as may be needed by a driver wishing to keep his or her eyes on the road instead of looking at a telephone device.

Speech recognition is employed in a mobile telephone to recognize a phrase, word, sound (generally referred to herein as utterances) spoken by the telephone's user. Speech recognition is therefore sometimes used in phonebook applications. In one example, a telephone responds to a recognized spoken name with an audible confirmation, rendered through the telephone's speaker output. The user accepts or rejects the telephone's recognition result on hearing the playback.

In human speech, each utterance has certain qualities that can be quantified, called prosodic parameters, which determine what the utterance sounds like. These are usually considered pitch or tone, timing of elements of the speech, and stress, usually represented as energy. Speech recognition systems use other features of speech, such as vocal tract shape, which are non-prosodic but help determine what was said. Human listeners are adept at discerning qualities of speech based in part on the prosodic parameters of the speech. Also, human speakers use prosody in speech to aid overall communication and to distinguish their speech from that of other speakers. Humans are thus naturally sensitive to prosody, and can easily determine the difference between "real" human speech and "synthesized" speech produced by a machine (speech synthesizer). In fact, synthesized speech using poor prosodic rules can be unintelligible to the human ear.

SUMMARY

Generally, aspects of the present invention feature methods and systems for synthesizing audible phrases (words) that include capturing a spoken utterance, which may be a word, and extracting both prosodic and non-prosodic information (parameters) there from, recognizing the word, and then applying the prosodic parameters to a synthesized (nominal) version of the word to produce a prosodic mimic phrase corresponding to the spoken utterance and the nominal word.

One aspect of the present invention features a method for speech synthesis, including receiving a spoken utterance; extracting one or more prosodic parameters from the spoken utterance; decoding the spoken utterance to provide a recognized word; synthesizing a nominal word corresponding to the recognized word; and generating a prosodic mimic word using the nominal word and the prosodic parameters.

Another aspect of the invention features a system for speech synthesis, including an audio input device that receives a spoken utterance; a pitch detector that detects a pitch of the spoken utterance; a signal processor that determines a prosodic parameter of the spoken utterance; a decoder that recognizes the spoken utterance and provides a corresponding recognized word; a speech synthesizer that synthesizes a nominal word corresponding to the recognized word; and a prosodic mimic generator that receives the nominal word and the prosodic parameter and generates a prosodic mimic word.

Yet another aspect of the present invention features a computer readable medium having stored instructions adapted for execution on a processor, including instructions for receiving a spoken utterance; instructions for extracting one or more prosodic parameters from the spoken utterance; instructions for decoding the spoken utterance to provide a recognized word; instructions for synthesizing a nominal word corresponding to the recognized word; and instructions for generating a prosodic mimic word using the nominal word and the prosodic parameters.

These and other aspects of the invention provide improved speech synthesis, especially in small mobile devices such as mobile telephones with voice activated commands and user interfaces. In one respect, better synthesis of audible confirmation messages is enabled, the audible confirmation messages having prosodic attributes resembling those of the user. Better speech synthesis sounds more natural and is more understandable to humans, therefore the present invention improves the usefulness and intelligibility of audible user interfaces.

Various features and advantages of the invention will be apparent from the following description and claims.

BRIEF DESCRIPTION OF THE DRAWINGS

For a fuller understanding of the nature and objects of the present invention, reference should be made to the following detailed description taken in connection with the accompanying drawings in which the same reference numerals are used to indicate the same or similar parts wherein:

Figure 1 is a block diagram of a mobile telephone device with a speech interface system;

Figure 2 is a block diagram of a process for synthesizing speech using a whole-word model; and

Figure 3 is a block diagram of a process for synthesizing speech using a phone-level model.

DETAILED DESCRIPTION

As discussed briefly above, human speech includes not only the substantive content (what words and sounds are made), but also information about the way the words and sounds are produced. Generally, a set of parameters (prosodic parameters) at least partially describes how a

spoken word or utterance is vocalized and what it sounds like. Examples of prosodic parameters are pitch, energy and timing. Better use of prosodic content can produce more natural and intelligible synthetic speech, a feature useful in modern communication systems like mobile telephones, which use synthesized audio interfaces.

A telephone device according to the present invention uses a speech synthesis circuit, logic, and executable code instructions to produce an audible signal delivered through its speaker output. By extracting and using prosodic features of a user's spoken words to synthesize and generate an audible output, the telephone device synthesizes high quality realistic-sounding speech that sounds like the user's voice. One specific application is in improving the quality and intelligibility of synthesized voice messages used to confirm spoken commands of a mobile telephone user.

Figure 1 is a block diagram of a mobile telephone device 10 having a voice user interface. The system includes input, output, processing, and storage components.

An audio input device 1000 receives a spoken utterance. The audio input device is a microphone, and more specifically, is the same microphone used to communicate over the mobile telephone device 10.

The audio input device 1000 provides the received audio input signal to a pitch detector 2100 and a Mel Frequency Cepstral Compact (MFCC) signal processor 2200, which extractsboth prosodic and non-prosodic parameter information from the received audio signal.

Decoder/speech recognition engine 2300 recognizes the spoken utterance and provides a recognized word to a speech synthesizer 2400. The recognized word is also provided as text to a visual display device (not shown).

The speech synthesizer 2400 synthesizes a nominal (default) form of the recognized word using rules that are pre-programmed into the system and that do not depend on the prosodic parameters of the spoken utterance.

To generate the prosodic mimic word, the prosodic mimic generator 2600 acts on the nominal synthesized word and applies the pitch, timing, or other prosodic parameters to the nominal synthesized word. The prosodic mimic generator 2600 adjusts the generated prosodic mimic word length by stretching or compressing the word in time. In the whole-word model of Figure 2, the beginning and end of the whole word act as temporal reference points, whereas in the phone-level model the individual phones act as the temporal reference points.

Once the prosodic mimic phrase is generated it is converted into a form suitable for audible output. The audio converter 2700 receives the prosodic mimic phrase and performs the necessary conversion to an electrical signal to be played by the audio output device 2800.

The embodiment shown in Figure 1 implements all but the input/output and memory storage components in a processor 20. Of course, more than one processor can be employed to achieve the same result. This includes embodiments employing multiple specialty processors, such as digital signal processors (DSPs).

Storage device 30 is a memory component that includes a machine-readable medium holding programmed software instructions. The machine is a data processor that reads and processes the instructions. The instructions are executed in the processor 20 or its components to carry out the functions of the system. An operating system is installed on the system that facilitates execution of the stored instructions for carrying out the voice recognition, processing, prosodic parameter extraction, speech synthesis, and mimic word generation. The storage device 30, is shared by the software instructions described herein, as well as by other program instructions belonging to other programs. For example, program instructions for controlling the ring tone, display graphics, and other features of the mobile telephone device can also reside in memory space allocated for these instructions within storage device 30.

Figure 2 is a block diagram of a process for generating synthesized utterances by using prosodic information from received spoken words. The functional blocks of the diagram correspond to physical components, as shown in Figure 1, which carry out the functions of the functional blocks. An utterance is divided into frames. The length of the frames affects the quality of the speech synthesis. The embodiment shown in Figure 2 processes utterances on a frame-by-frame basis, where a frame is a predefined time segment. For speech applications, a frame length that is too long can lead to inaccuracies and low quality speech synthesis, while a frame length that is too short requires more computing resources (processing, storage, etc.). In the described embodiment, the frame length is approximately 10-20 milliseconds in duration.

An input device, such as a microphone, captures a spoken utterance 102 (for example, the phrase "CALL HOME") at step 100. The spoken utterance 102 corresponds to an action to be taken by the mobile telephone device, here calling the user's home phone. In this example, the telephone looks up and dials the telephone number (HOME) whose name was spoken.

The system analyzes spoken utterance 102 for its prosodic parameters and extracts the values for the prosodic parameters. The system extracts, for example, the pitch of the spoken utterance. Pitch generally refers to the overall frequency content of the voice. Step 110 depicts pitch detection.

The system also extracts the spectral content, e.g., mel cepstra, and energy content of spoken utterance 102 at step 120. A MFCC analyzer measures the MFCC Cepstrum of the spoken utterance 102. The MFCC analyzer outputs frames of prosodic parameters at step 122..

A decoder or speech recognition engine decodes or recognizes the spoken utterance at step 130. The decoder employs hardware and software to select a recognized word from a set of possible known words. The decoder recognizes a recognized word, corresponding to the spoken utterance, and provides the word as a text output 132 to visually indicate the results of the decoding. A display device of the mobile telephone shows the text output 132 to the user.

The decoder also delivers the recognized word 134 to a speech synthesizer that uses the recognized word and a set of default programmed (nominal) synthesis rules to generate synthesized nominal word frames at step 140. In this embodiment, the decoder uses a whole-word model , and the synthesis takes place at the word level.

A prosodic mimic generator generates the prosodic mimic phrase using the recognized word's nominal synthesized frames 142, the captured prosodic parameters provided in the pitch per frame 112 and the actual frames 124. The prosodic mimic generator applies the prosodic parameters to the nominal frames 142 on a frame-by-frame basis. Furthermore, in step 150, the prosodic mimic generator temporally aligns the generated mimic word with the nominal word, at a whole-word level. In other words, the recognized word 134 is aligned in time with the corresponding captured spoken word by forcing the start and end points of the nominal word to correspond to those of the spoken word.

The prosodic mimic generator applies the captured prosodic parameters, such as pitch, to the nominal word, thereby mimicking the prosody of the spoken utterance 102. The prosodic mimic generator also adjusts the length of the generated phrase by stretching and compressing the phrase to obtain the desired length. Stretching and compression of the prosodic mimic phrase is done by adding and removing frames, respectively, from the phrase in order to match the phrase length to that of the spoken utterance. The result is a synthesized prosodic mimic phrase that, owing to its prosody, mimics the original spoken word in its content and its sound.

An audio converter receives the generated prosodic mimic phrase and converts the nominal frames with the applied actual timing and pitch 152 into an audio signal to be played on the mobile telephone's speaker (step 160). The speaker is the same speaker over which the user hears the ordinary telephone communication output.

The end result of the process described above is a natural-sounding audible phrase resembling the originally spoken utterance 102. This synthesized mimic phrase is used as an audible confirmation message played back to the mobile telephone user to confirm the command to be carried out or the name to be dialed.

Figure 3 illustrates a process using a phone-level model, according to which words are synthesized at a finer level of detail than is done in the whole-word model. Generally, phones are acoustic constituents of speech. A spoken language includes a set of phones which are used to form the sounds of the spoken language. For example, "HOME" contains three phones: "H", "O" and "M." It is possible to improve the quality and accuracy of speech synthesis if speech is treated at the phone level rather than on a whole-word level.

An input device, such as a microphone, captures a spoken utterance in step 100, as described earlier. One or more signal processors and a pitch detector extract prosodic parameters (pitch, energy and/or timing) from the spoken utterance 102. The pitch detector detects the spoken utterance's pitch at step 110, and a MFCC analyzer extracts the mel cepstra and timing parameters at step 220. Some of the timing information may come from a decoder, which may be part of a speech recognition system.

A decoder recognizes the speech at step 230. The decoder outputs a selected recognized word 232 to a visual display unit, and also outputs individual phones 234 and alignment information of the recognized word to a phonetic speech synthesizer. The decoder provides alignment information 236 for use in generating a prosodic mimic phrase later.

A phonetic speech synthesizer takes the phones and alignment output from the decoding step 230 and performs a phone-level synthesis of the recognized words at step 240. The speech synthesizer outputs frames from the phonetic synthesis 242.

Parameter lookup step 250 is based on nominal frame phones, and provides nominal frames and nominal alignment information 252.

A prosodic mimic generator receives the nominal frames at step 260, as well as the captured actual frames 224, alignment information 236, pitch-per-frame data 212, and the

nominal frames with nominal alignment 252. The prosodic mimic generator outputs a set of nominal frames having timing, energy and pitch derived from the input spoken phrase 102. This is the prosodic mimic phrase 262.

As described for the earlier embodiment of Figure 2, the nominal selection is synthesized using the extracted prosodic parameters obtained from the spoken word. However, in this embodiment, rather than time-aligning the nominal word to the spoken word, the constituent phones are used as the temporal indexing points or boundary markers that delineate the nominal-to-spoken alignment process. In other words, the embodiment of Figure 3 aligns the phones within words, as well as the words themselves, thereby imposing greater constraints on the overall time-alignment process.

As described previously, an audio converter converts the prosodic mimic word 262 to an audio signal in step 270. An audio output device delivers an audible signal to the telephone's user at step 280. A digital-to-analog converter converts the digital prosodic mimic word signal into a signal that can be played on the telephone device's speaker.

The concepts described above are not limited to the uses recited in the illustrative embodiments provided, but can be extended to other systems and circumstances. For example, the application of such techniques and devices can extend to any voice-driven electronic device, including personal planners, toys, automotive navigation equipment, home electronics, home appliances, and computing devices in general.

The present system and methods are also not limited to words only, but to any portion of a word or combination of words, phrases, sentences, audible gestures, etc. in any spoken language. Therefore, we refer to any and all of these as utterances.

These concepts may be used in combination with other human-machine interfaces. For example, not only does the mobile telephone provide its user with audible and/or visual feedback to confirm a command or number to be dialed, but it can also require actions on the part of the user to accomplish such commands. The user may be required to press a confirmatory button on the mobile telephone to indicate agreement with the recognized and synthesized word, or the user may be required to say "YES" or "OK" to make a final acceptance of a synthesized audible message.

Upon review of the present description and embodiments, those skilled in the art will understand that modifications and equivalent substitutions may be performed in carrying out the

invention without departing from the essence of the invention. Thus, the invention is not meant to be limited by the embodiments described explicitly above, rather it should be construed by the scope of the claims that follow.

What is claimed is: